APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Nov. 2006, p. 7286–7293 Vol. 72, No. 11
0099-2240/06/$08.00+0 doi:10.1128/AEM.01398-06

Toward a More Robust Assessment of Intraspecies Diversity, Using Fewer Genetic Markers▽†

Konstantinos T. Konstantinidis,*¶ Alban Ramette,§¶ and James M. Tiedje

Center for Microbial Ecology, Michigan State University, East Lansing, Michigan

Received 16 June 2006/Accepted 5 September 2006

Phylogenetic sequence analysis of single or multiple genes has dominated the study and census of the genetic diversity among closely related bacteria. It remains unclear, however, how the results based on a few genes in the genome correlate with whole-genome-based relatedness and what genes (if any) best reflect whole-genome-level relatedness and hence should be preferentially used to economize on cost and to improve accuracy. We show here that phylogenies of closely related organisms based on the average nucleotide identity (ANI) of their shared genes correspond accurately to phylogenies based on state-of-the-art analysis of their whole-genome sequences. We use ANI to evaluate the phylogenetic robustness of every gene in the genome and show that almost all core genes, regardless of their functions and positions in the genome, offer robust phylogenetic reconstruction among strains that show 80 to 95% ANI (16S rRNA identity, >98.5%). Lack of elapsed time and, to a lesser extent, horizontal transfer and recombination make the selection of genes more critical for applications that target the intraspecies level, i.e., strains that show >95% ANI according to current standards. A much more accurate phylogeny for the *Escherichia coli* group was obtained based on just three best-performing genes according to our analysis compared to the concatenated alignment of eight genes that are commonly employed for phylogenetic purposes in this group. Our results are reproducible within the *Salmonella*, *Burkholderia*, and *Shewanella* groups and therefore are expected to have general applicability for microevolution studies, including metagenomic surveys.

Understanding the extent and importance of the genetic and biochemical diversity among strains of the same or very closely related species is a cornerstone issue for many microbiological disciplines, such as taxonomy, diagnosis, epidemiological studies, (environmental) diversity surveys, biogeography, etc. The multilocus sequence typing (MLST) method has recently emerged as the method of choice for exploring and cataloguing intraspecies genetic diversity (4, 6, 14), thus setting the stage for linking the genetic diversity to the biochemical and functional diversities of species. Typical applications of the MLST method employ the sequencing of six to eight genes (loci) and subsequent phylogenetic analysis of the concatenate sequence alignments to reveal the exact genetic relationships among the strains analyzed (5, 6, 14).

Although the advantages of the MLST method over several traditional methods for strain genotyping are now well documented in the literature (4, 6), several issues remain less clear. Most importantly, it remains unknown how well the phylogenetic relationship based on eight loci, which are selected primarily based on being unlinked in the genome (e.g., to avoid hitchhiking of selection and recombination events) and offer conserved sites for PCR primer design (14) rather than on their phylogenetic robustness, approximates the real phylogenies of the strains studied. It is also largely uninvestigated how comparable the results derived by different sets of genes are and which functional classes of genes provide better phylogenetic markers. Last, it remains unclear whether a smaller number of loci, which would substantially economize on sequencing cost and time, could give results comparable to those based on eight loci. The principal reason that these issues remain largely uninvestigated is the lack of a robust and highly accurate method or measurement that can be used as a reference standard to compare the phylogenetic informativeness of every gene in the genome and hence identify the best-performing genes for phylogenetic purposes. Even if such a measurement were available, however, the exact methodology of how to perform the analysis at the whole-genome level would remain challenging.

The recent availability of genomic sequences for a number of closely related bacterial strains has made it possible for the first time to construct highly accurate phylogenetic reconstructions among the strains (3, 11). Such reconstructions may be used as a reference standard to provide new insights into the issues described previously. Toward these goals, we have analyzed four important bacterial groups that are currently best represented with genomic sequences, and these genomic sequences sample various levels of resolution within and between closely related species. Using state-of-the-art methods for phylogenetic analysis, we built whole-genome-based phylogenies for the members of a group and used them to investigate the performance of every gene in the genome for phylogenetic purposes. A much more accurate phylogenetic reconstruction was achieved by using just three of the best-performing genes identified by our analyses compared to classical MLST approaches that employ six to eight genes. Our results are ex-

* Corresponding author. Present address: 15 Vassar Street, Room 48-336, Massachusetts Institute of Technology, Cambridge, MA 02139. Phone: (617) 253-3897. Fax: (617) 253-2679. E-mail: konstan1@mit.edu.

† Supplemental material for this article may be found at http://aem.asm.org/.

§ Present address: Celsiusstrasse 1, Max Planck Institute for Marine Microbiology, Bremen 28359, Germany.

¶ K.T.K. and A.R. contributed equally to this work.

▽ Published ahead of print on 15 September 2006.

pected to have important practical implications for how strain genotyping and phylogenetic analysis are performed, particularly among closely related organisms.

MATERIALS AND METHODS

Bacterial genomes used in this study. Four bacterial groups, namely, *Escherichia coli*, *Salmonella* spp., *Shewanella* spp., and *Burkholderia* spp., which had 12, 11, 9, and 11 sequenced representatives, respectively, were included in the analyses. The genomic sequences and sequence annotations for 15 of the 43 genomes, which were published at the time of this study (January 2006), were obtained from NCBI's FTP site at ftp://ftp.ncbi.nih.gov/. The remaining 28 genomes were at various stages of the gap-closing phase and typically had fewer than 20 gaps in their sequences. These genomes were *Salmonella bongori* 12419, *Salmonella enterica* serovar Enteritidis PT4, *Salmonella enterica* serovar Gallinarum 287/91, *Salmonella enterica* serovar Typhimurium DT104, *Salmonella enterica* serovar Typhimurium DT2, *Salmonella enterica* serovar Typhimurium SL1344, *Shigella dysenteriae* M131649, *Shigella sonnei* 53G, *Escherichia coli* 042, *Escherichia* coli E2348/69, and *Burkholderia cenocepacia* J2315, produced by the Sanger Center and obtained through the Sanger FTP site at ftp://ftp.sanger.ac.uk/pub/; *Escherichia coli* HS and *Escherichia coli* E24377A, produced by The Institute for Genomic Research (TIGR) and obtained through their website at http://www.tigr.org; and *Shewanella putrefaciens* CN-32, *Shewanella putrefaciens* ANA-3, *Shewanella putrefaciens* W3-18-1, *Shewanella baltica* OS1155, *Shewanella* sp. MR-4, *Shewanella* sp. MR-7, *Burkholderia cenocepacia* AU1054, *Burkholderia cenocepacia* HI2424, *Burkholderia ambifaria* AMMD4, *Burkholderia* sp. 383, *Burkholderia vietnamiensis* G4, and *Burkholderia xenovorans* LB400, produced by the Joint Genome Institute (JGI) and obtained through their website at http://www.jgi.doe.gov. The genomic scaffolds corresponding to two distinct *Shewanella* sp. and one *Burkholderia* sp. population recovered in the shotgun sequencing of the Sargasso Sea (22) were also included in the analyses and obtained from NCBI's FTP site.

Conserved gene cores. For each group, the conserved gene core, i.e., the genes that are shared by all members of the group, was determined using the following strategy. All annotated genes in one of the published genomes of the group (hereafter referred to as the reference genome for the group) were searched against the genomic sequences of the remaining genomes of the group by using the BLASTn (nucleotide search) algorithm, release 2.2.9 (2). The best match, when it showed better than 50% identity over at least 70% of the length of the gene in the reference genome, was extracted from the genomic sequence with a custom PERL script and searched back (BLASTn) to the reference genome's genes to identify the reciprocally best-match-conserved (and presumably orthologous) gene set. The genes of the reference genome that were reciprocally best match conserved in all genomes of the group constituted the conserved gene core for the group. The previous strategy circumvented the problem of inconsistencies in the annotations of the published genomes and the need for annotating the draft genomes. The BLASTn algorithm was run with the following settings: $X = 150$ (drop-off value for gapped alignment), $q = -1$ (penalty for nucleotide mismatch), and $F = F$ (filter for repeated sequences); the rest of the parameters were at default settings. These settings give better sensitivity with moderately diverged sequences than default settings, which target highly identical sequences (10). Because the groups encompass closely related genomes (all members show >80% average nucleotide identity among themselves), using a lower cutoff or searching at the amino acid level (as opposed to the nucleotide level) within a group did not substantially differentiate the conserved gene core for the group. Genes conserved between groups were determined using an identical strategy but searching at the amino acid level (BLASTp) and using a cutoff of 30% amino acid identity over at least 70% of the length of the gene, to accommodate the evolutionary distances between the groups, which were greater than the intragroup distances (10).

Phylogenetic gene analysis. For every reference gene in the conserved gene core of a group, an alignment of all of its orthologs within the group (therefore, the number of orthologs equals the number of genomes in the group) was built using ClustalW software (21). MODELTEST version 3.7 software (17) in combination with PAUP (20) was used to find the most plausible evolutionary model (out of 56 models in total) for each gene via the Akaike information criterion test (16, 17). The best model was subsequently used in a maximum-likelihood (ML) analysis as implemented in the PAUP software (20) to build a phylogenetic tree and calculate the ML-based distances between all pairs of genomes in a particular group. For instance, for the *E. coli* group that includes 12 genomes, there are 66 nonredundant pairs of orthologs (equal to the number of pairs of *E. coli* genomes) for every gene of the 2,646 core genes in all 12 *E. coli* genomes. A whole-genome analysis based on the concatenated alignments of all core genes of a group was also carried out using a strategy identical to the one described above for individual genes. The gene-based ML distances were compared to the whole-genome-based ML and the average-nucleotide-identity (ANI) distances (10) for the same pairs of genomes using the nonparametric Kendall τ correlation as implemented in the SPSS Grad Package (18). The latter was used because neither ML nor ANI values were normally distributed. Average-nucleotide-identity values were calculated based on the nucleotide level identities, as computed by the BLASTn algorithm, of only the core genes of a group, i.e., the same genes used in the whole-genome-based ML analysis (hereafter referred to as ANIo), as opposed to the whole genome (hereafter referred to as ANIg to avoid confusion) in our previous study (10). Custom PERL scripts were used to semiautomate the process and to parse the outputs of the software as needed.

Statistical analyses. The statistical significance of the correlation between the ML distances based on individual genes and the ANIo (or whole-genome ML) distances was evaluated for the *E. coli* group by using a delete-half jackknife approach. For every gene, a random selection of 33 pairs of genomes (of the total 66 nonredundant pairs in the *E. coli* group) was made without replacement and the ML gene-based distances for these 33 pairs were compared to the ANIo values for the same pairs of genomes by using the Kendall τ correlation test. The procedure was repeated 1,000 times, and the distribution of the Kendall τ values was used to define the 2.5 and 97.5% quantiles, which define the lower and upper limits of the 95% confidence interval for the Kendall τ values, respectively, as shown in Fig. 2.

To estimate how well the average distances for a random selection of a given number of genes correlated to the ANIo (or whole-genome ML) distances, the following procedure was applied to the *E. coli* group. A random selection without replacement of a given number of genes was made, and the average ML distances among all 66 pairs of genomes in the *E. coli* group based on all genes selected were calculated and compared to the ANIo values for the same genome pairs by using Kendall's τ correlation test. The procedure was repeated 1,000 times, and the distribution of the Kendall τ values was used to define the 95% confidence interval for the Kendall τ values, as shown in Fig. 3. Figure S1 in the supplemental material represents a graphical description of the methodology described above for calculating Kendall τ values for individual genes and their confidence intervals based on bootstrap analysis.

The randomness of the distribution of Kendall τ values along the mean gene position in a given genome was determined by the Runs test for a significance level (P) of <0.05, their autocorrelation by a Moran I correlogram (significance determined by 1,000 bootstrap replicates for each distance class), and their periodicity by a Lomb periodogram (13). Distance classes were made so that each consisted of an equal number of observations (to make statistical comparisons meaningful), and the first class spanned a distance of about 100 consecutive genes. All statistical calculations were implemented with the R statistical package (http://cran.r-project.org/).

RESULTS

Diversity within each group and the robustness of the ANI measurement. The four bacterial groups chosen have contrasting ecologies, e.g., environmental (*Burkholderia* and *Shewanella*) versus pathogenic (*E. coli* and *Salmonella*), and genome sizes, ranging from 4 Mb (*Shewanella*) up to 9 Mb (*Burkholderia*). Furthermore, these groups encompass various levels of relatedness, with *E. coli* and *Salmonella* being very close relatives, *Shewanella* being a more distant relative of the former two within the same phylum (γ-proteobacteria), and *Burkholderia* being a member of the β-proteobacterial phylum. Therefore, the robustness of the conclusions obtained for one group can be tested against that for increasingly more-distant groups.

First, an ML analysis of the concatenated conserved gene core of each group (at least 1,400 genes aligned) was performed to determine the phylogenies of the genomes in a group, and the ML-based distances among the genomes were calculated. The ML analysis employed the best evolutionary model (see Materials and Methods), and thus, it represented a particularly powerful measurement of the evolutionary distances between the genomes in a group. Interestingly, the best

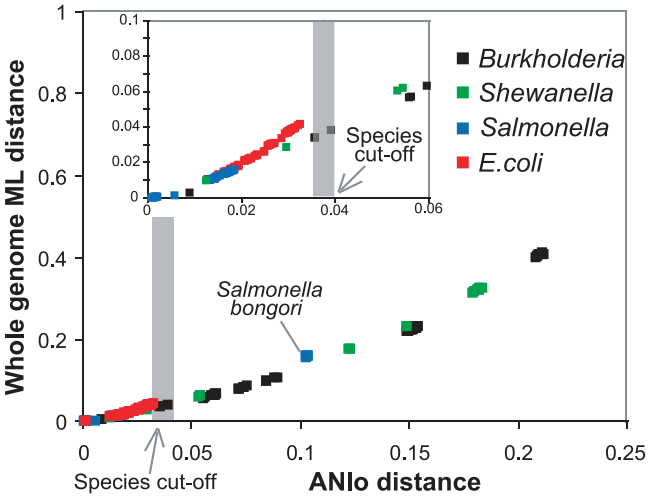

FIG. 1. Genetic diversity within each of the four bacterial groups studied, based on the ML and ANIo measurements. Each square represents a pair of genomes from one group, colored according to the group to which the genomes belong (see legend). Whole-genome ML distances between two genomes in the pair (*y* axis) are plotted against their ANIo distances (*x* axis). The gray area corresponds to the current species cutoff for bacteria.

model for sequence evolution was the same for all groups, i.e., the time-reversible model with six categories of sites and gamma distribution of variation rates among sites (i.e., GTR+I+G). Only the values for individual variables of the model were different among groups. In contrast, the best models for individual genes were more variable among genes (data not shown but available upon request).

The whole-genome-based ML distances between all pairs of genomes in a group were compared to the ANIo distances for the same pairs of genomes. A perfect correspondence between the two measurements was observed for all four groups, i.e., r^2 values of >0.98 for all groups (Fig. 1). It also appeared that the relationship was linear for ANIo distances up to 0.1 to 0.15 (i.e., 85 to 90% identity). Beyond this area, multiple substitutions at the same site, which are not considered in the ANIo measurement (but are considered in the ML analysis), presumably made the relationship nonlinear (but equally strong). In other words, the robustness and discriminative power of the ANIo measurement remain equally strong below 85% identity but the absolute evolutionary distance becomes gradually and uniformly more compressed in the units of ANIo below the 85% identity level. In any case, however, these results suggest that ANIo distances can be used interchangeably with whole-genome ML distances for short evolutionary scales, which is also consistent with our previous study (10). Furthermore, ANIo distances are easier to realize conceptually than ML-based distances; therefore, for the procedures indicated in the remaining text, we used ANIo measurement unless otherwise noted. Finally, because the conserved gene core in a group is enriched in housekeeping genes, which tend to show higher sequence conservation than the genome average, the 70% DNA-DNA reassociation hybridization (DDH) standard for species demarcation in bacteria (19, 23) corresponded to ~96% ANIo in the current analysis as opposed to 94 to 95% ANIg in our previous study using the whole-genome sequences (10).

The four groups studied encompass different levels of genetic diversity. The *Salmonella* group, with the exception of the *S. bongori* genome, includes very closely related genomes (i.e., showing more than ~98% ANIo among themselves) that should all belong to the same species according to the current standard (10, 23); the *E. coli* group, similar to the *Salmonella* group, includes genomes that should belong to the same species but are slightly more diverse than the *Salmonella* ones, i.e., most genomes show 96 to 98% ANIo among themselves (Fig. 1). The *Shewanella* and *Burkholderia* groups include genomes that belong to closely related species, i.e., showing 80 to 95% ANIo among themselves, in addition to a few genomes that belong to the same species (Fig. 1). Therefore, these four groups comprise at least three different levels of resolution among closely related organisms. No genome that showed <80% ANIo to another member of the same group was included in the analysis, because our focus was species-level differences, while all members of a group show >98.5% small-subunit rRNA gene identity among themselves (<98.5% small-subunit rRNA between groups) (data not shown).

Evaluation of the phylogenetic robustness of every gene in the genome. The performance of individual core genes relative to that of the whole genome was evaluated at two different levels: (i) distance matrix comparison, in which the gene-based ML distances among all pairs of genomes in a group were compared to the ANIo distances for the same pairs of genomes to identify which genes are good predictors of whole-genome-based distances and are thus good candidates for phylogenetic analysis within the groups, and (ii) tree comparison, in which a maximum-likelihood Kishino-Hasegawa (KH) test (9), including branch lengths, was employed to test whether the gene-based phylogeny was statistically different from the whole-genome-based phylogeny based on the gene alignment. The consistency index and retention index as implemented in PAUP (20) were also computed to test whether nonsignificant *P* values in the KH test were merely the result of phylogenetic noise as opposed to a strong phylogenetic signal. In general, there was a good correlation between showing a strong correlation with ANIo and having a low *P* value in the KH test or a high consistency index (data not shown).

Our results revealed a very interesting trend: when the analysis included genomes from different species, i.e., covering the whole range of evolutionary relatedness studied here (80 to 100% ANI), most genes in the genome showed very strong correlations with ANIo (see, for example, the *Burkholderia* group in Fig. 2B). In other words, robust phylogenetic reconstruction of well-separated (i.e., not too closely related) strains is feasible for almost any of the core genes. When the analysis included only closely related strains of the same species, as in the *E. coli* group, then many genes in the genome showed poor correlations with ANIo (Fig. 2C). The trend is nicely exemplified in the *Salmonella* group, where the removal of the *S. bongori* genome, the only genome that is more distantly related to the remaining *Salmonella* genomes, has a dramatic effect on the distribution of the Kendall τ correlation coefficients for the core genes of *Salmonella* (Fig. 2A). The correlations observed for individual genes appear to be independent of the genomes used. Indeed, when a jackknife analysis, using half of the

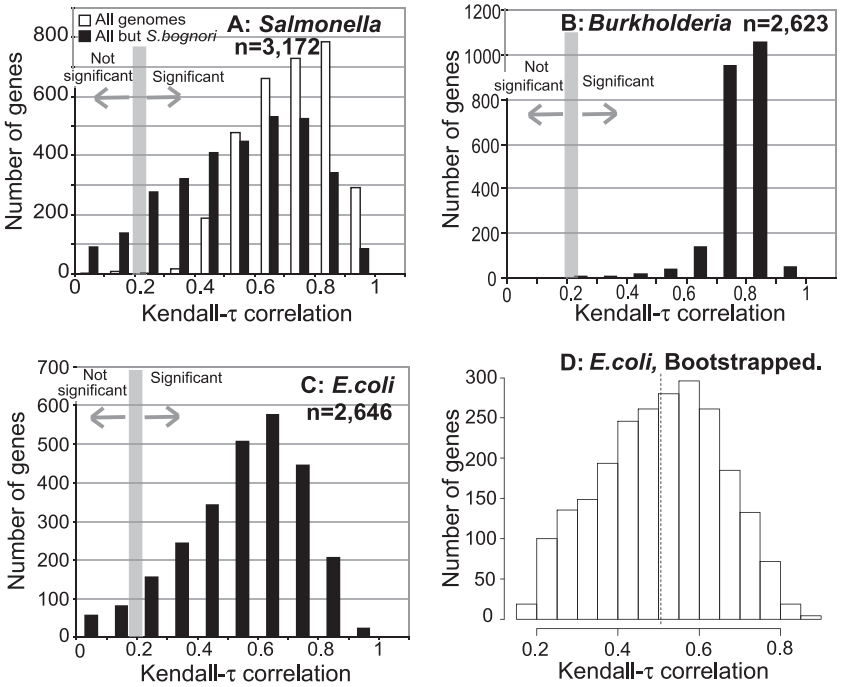

FIG. 2. Performance of individual genes against the whole-genome average. The individual-gene-based distance matrix for all genome pairs in a group was compared to the ANIo matrix for the same genome pairs by using the nonparametric Kendall τ correlation test. Graphs show the distribution of the Kendall τ values for all core genes within a group (1, perfect correlation; 0, no correlation). The area that corresponds to a significant correlation ($P < 0.05$) is also designated. The *E. coli* (A), *Salmonella* (B), and *Burkholderia* (C) groups are shown. Panel D shows the distribution of the Kendall τ values for the genes in the *E. coli* group, which were significant, as determined by the bootstrap approach (see Materials and Methods for details). n, number of genes.

genome pairs and 1,000 replicates, was performed for every gene in the *E. coli* group, the range of Kendall τ correlation coefficients was within a 13% difference from the correlation coefficient observed with all genome pairs in 95% of the replicates. Last, there was weak ($R^2 = 0.11$ for the *E. coli* group) but significant ($P < 0.001$) correlation between the Kendall τ correlation coefficient and the length of the gene. Data for the top 20 genes in terms of correlation with ANIo for each group considered as well as for selected genes that have been used for MLST studies in any of the four groups are shown in the supplemental material and are also available through the Ribosomal Database Project website (http://rdp.cme.msu.edu/). Data for all genes within a group are available from the authors upon request.

Using a small number of genes to predict whole-genome relatedness. In typical MLST applications, six to eight genes are sequenced and phylogenetic analysis of their concatenated alignment is performed to reveal the "average" phylogeny. To investigate how well the average for a given number of randomly selected genes correlates with (and thus could predict) the ANIo and what the optimum number (if any) of genes for predicting whole-genome relatedness is, a resampling strategy without replacement was performed for the *E. coli* group (see Materials and Methods for details). Our results showed that the average distances for even two randomly selected genes almost always showed a significant correlation (Kendall $\tau > 0.221$, $P < 0.05$) with ANIo distances. This represents the worst-case scenario, where genes with poor correlation with ANIo were included in the selection (Fig. 3, solid squares). Conversely, if the two genes were selected to be among the best-performing genes in terms of correlation with ANIo, then the correlation of their average value with the ANIo gave an r^2 value of ~0.81 (Fig. 3, open squares). The analysis also showed that the higher the number of genes sampled, the higher the

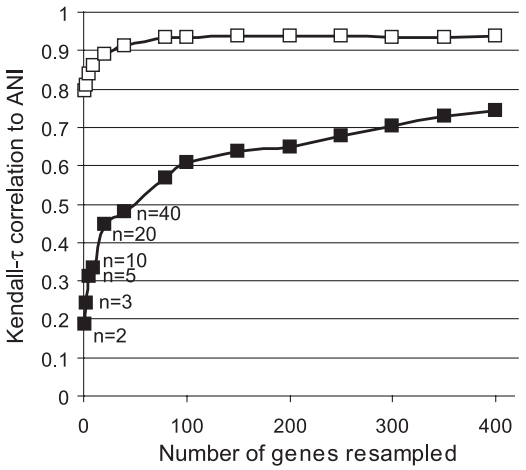

FIG. 3. Upper and lower confidence levels for the Kendall τ correlation coefficients for a given number of genes. The upper (open squares) and the lower (solid squares) 95% confidence levels for the Kendall τ correlation coefficients between the averages for the ML distances for a given number of genes (*x* axis) and the ANIo distances are shown for the *E. coli* group. See Materials and Methods for details on the calculation of the confidence intervals.

TABLE 1. Analysis of genes that have been used in MLST studies for *E. coli* (top eight) and the best-performing genes identified in this study (lower three)[a]

GenBank accession no.	Gene	Source or reference[b]	Annotation[c]	COG[d]	Kendall[e]	Kendall-20[f]	HI[g]	RI[h]	KH *P*[i]	No. of sites[j]	Total no. of sites[k]
NP310589	*zwk*	1	Glucose phosphate dehydrogenase	G	0.599	0.612	0.19	0.79	0.018	36	1,477
NP311583	*recA*	MLST-DB	Recombinase A	L	0.696	0.701	0.25	0.72	0.16	14	1,062
NP308554	*adk*	MLST-DB	Adenylate kinase	F	0.334	0.51	0.32	0.61	0.016	21	645
NP311390	*ppk*	1	Polyphosphate kinase	P	0.404	0.523	0.34	0.49	0.002	44	2,067
NP310344	*fumC*	MLST-DB	Fumarate hydratase	C	0.562	0.594	0.17	0.71	0.073	37	1,404
NP309635	*icdA*	MLST-DB	Isocitrate dehydrogenase	C	0.144	0.274	0.45	0.52	0.027	53	1,251
NP312174	*aroE*	MLST-DB	Shikimate 5-dehydrogenase	E	0.333	0.476	0.46	0.58	0.001	36	819
NP312136	*mdh*	MLST-DB	Malate dehydrogenase	C	0.322	0.417	0.25	0.71	0.173	23	939
NP313063	*tyrB*	This study	Aspartate aminotransferase	E	0.835	0.847	0.21	0.76	0	47	1,194
NP309178	*torC*	This study	Trimethylamine *N*-oxide reductase	C	0.816	0.794	0.24	0.75	0.022	40	1,173
NP311675	*gudX*	This study	Putative glucarate dehydratase	MR	0.786	0.808	0.32	0.65	0.864	46	1,023

[a] KH tests compared the individual-gene and whole-genome ML trees based on individual-gene alignment. There are several more genes in the genome that perform comparably to the three genes shown here and were used in the phylogenetic analysis (Fig. 4C). Information for the top 20 genes is shown in the supplemental material. Information for all genes in the core of the *E. coli* group or for genes in any of the other three groups considered is available from the authors upon request.

[b] Reference numbers refer to the study that employed the respective genes for MLST. MLST-DB, the *E. coli* MLST database, Max Planck Institute, Germany (http://web.mpiib-berlin.mpg.de/mlst/dbs/Ecoli/documents/primersColi_html).

[c] The gene annotation from GenBank.

[d] The COG functional category of the gene.

[e] Kendall τ correlation with ANIo.

[f] Kendall τ correlation with ANIo for the 20-genome data set.

[g] Homoplasy index.

[h] Consistency index.

[i] *P* value for the KH test.

[j] The number of informative sites in the gene sequence.

[k] The total number of sites used in the ML analysis.

confidence level for the Kendall τ correlation between their average value and the ANIo, i.e., a more robust prediction of ANIo was more likely. The lower confidence levels appeared to substantially increase for up to about 100 genes, after which they increased only slightly with more genes sampled. For instance, a random selection of 10, 100, and 300 genes gave lower Kendall τ correlation coefficients of 0.33, 0.61, and 0.7, respectively (Fig. 3, solid squares).

In silico MLST evaluation. As exemplified previously by the *E. coli* and *Salmonella* groups, the selection of genes for MLST applications that target intraspecies diversity could be very critical because not all genes perform well at this level (Fig. 2), and thus, the accuracy may be very variable (Fig. 3). To further investigate this, we selected three of the best-performing genes for the *E. coli* group according to our analysis and built a phylogeny based on the concatenated alignment of the three genes by using full-length gene sequences. The three genes were selected primarily based on their correlations with ANIo (the primary criterion; see above), while the tree-based criteria (see above) were used as additional, secondary criteria to better refine the selection of the best genes. We then compared this phylogeny to the phylogeny built on eight genes that are frequently used in MLST studies for the *E. coli* species (1, 8, 15) and the whole-genome phylogeny (i.e., the reference phylogeny; all genes used and their statistics are summarized in Table 1). Our results revealed that with as few as three genes (3,300 bases in total), a phylogeny more congruent to the whole-genome-based phylogeny was achieved than what was found for the classical MLST application (8 genes; 9,500 bases in total). This was evident both in terms of tree topology, i.e., with KH test *P* values of 0.709 and 0.02 for the classical MLST tree and our top-three-gene tree, respectively, and in terms of branch length and bootstrap values (all nodes have >50% bootstrap support in the top-three-gene tree) (Fig. 4).

Functional and spatial evaluation of the best-performing genes. The functional annotations of the genes with significant correlations with ANIo were examined more closely to reveal whether genes that are good predictors of ANIo belong to specific functional categories. The functional annotations of the genes were extracted from the GenBank files for the reference genomes. An in-house annotation of the reference genomes, using the Cluster of Orthologous Genes database (COG) as described before (12), was also performed to look for major trends in our data, at the general functional-category level. The number of genes that were good predictors (i.e., showed significant Kendall τ correlation) of ANIo values did not depend on COG classification: each COG category had a constant proportion of genes that were good predictors of ANIo, and this was reproducible among the four groups considered (Fig. 5). Further, the genes with better correlation with ANIo were relatively evenly distributed throughout the genome, e.g., no systematic clustering toward the origin, the middle, or the terminus of replication was evident (Fig. 6, outer circle). Hence, no major trend or functional bias was evident. Some minor functional trends were evidenced, however, and are summarized below.

Almost all core genes in the more diverse groups, i.e., *Burkholderia* and *Shewanella*, showed significant correlations with

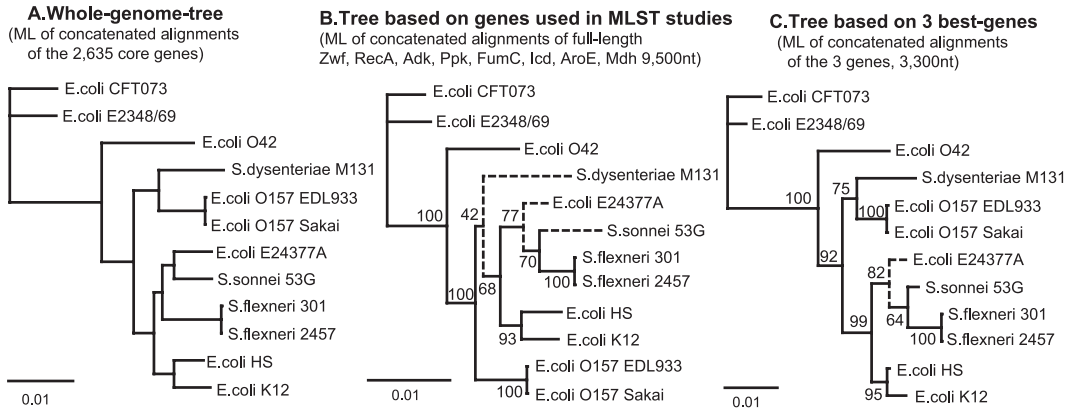

FIG. 4. Improved phylogenetic reconstruction in an MLST-like application, using only three of the genes in the genome. Three maximum-likelihood trees are shown, one based on the concatenated alignment of all 2,635 core genes for the *E. coli* group (A), one based on the concatenated alignment of 8 genes frequently used in MLST studies for the *E. coli* group (B), and one based on the concatenated alignment of 3 of the best-performing genes according to our analysis (C). Dashed branches designate the major differences between the trees in panels B and C and the whole-genome tree (A). nt, nucleotides.

ANIo (Fig. 2B), and thus, the proportion of genes that were good predictors of ANIo was almost invariable for every COG category at ~100%. For the *E. coli* group, however, this proportion was only 72%. Moreover, several COG categories, such as the informational categories (Fig. 5A, point J), which include the ribosomal proteins, polymerases, and tRNA synthetases, etc., as well as the categories of hypothetical (Fig. 5A, No COG) and conserved hypothetical (Fig. 5A, point S) proteins, had a higher proportion of "poor-predictor" genes. Conversely, metabolism and cellular-process categories (for example, Fig. 5A, points P, M, and C) had a higher proportion of "good-predictor" genes relative to the average for all categories (Fig. 5). The best-performing genes included, but were not limited to, dehydrogenases and transport enzymes, for example. The *Salmonella* group (without the *S. bongori* genome), on the other hand, which encompasses a diversity comparable to that of the *E. coli* group, seemed to have a more uniform proportion of "good-predictor" genes for every COG category than the *E. coli* group (Fig. 5B). These results indicate that there may be small but qualitatively significant differences in terms of which genes are the best predictors of ANIo, even among very closely related groups.

The distribution of the Kendall τ values in the genome of *E. coli* strain Sakai was strikingly nonrandom ($Z = -5.466$, $P \ll 0.001$; Runs test), but there was no systematic increase or decrease detected (Fig. 6A). A visual inspection of the distribution patterns revealed instead the existence of a wave-like structure, which was further statistically confirmed by analysis of the Moran *I* correlogram, which displayed a succession of significant correlation coefficients in peaks and troughs along increasing distance classes (Fig. 6B) (13). The Moran *I* correlogram also revealed that successive Kendall τ values were not independent of each other within the range of 80 to 100 consecutive genes (i.e., the first distance class) (Fig. 6B), i.e., that the data were significantly autocorrelated. Further statistical analyses were performed to reveal the periodicity of the distribution, and using the Lomb periodogram, we also identified the most significant period to be about 100 genes ($P < 0.01$), which was very consistent with the results obtained from the Moran *I* correlogram.

Future research including more genomes and phylogenetically different groups is needed to shed light on the underlying mechanisms leading to the observed periodicity in the robustness of the phylogenetic signals of individual genes. It does

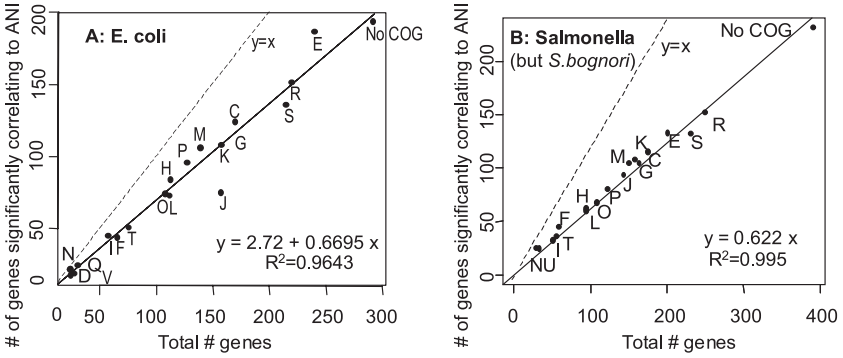

FIG. 5. Functional annotations of the genes with significant correlation with ANIo values. The number of genes with significant correlation with ANIo (Kendall τ correlation > 0.221, $P < 0.05$) (*y* axes) is plotted against the total number of genes in a COG functional category (*x* axes). Using a higher cutoff for Kendall τ correlation (e.g., >0.35) does not change the results shown. The letters on the graphs correspond to the COG individual functional categories according to the category designations on the COG website (http://www.ncbi.nlm.nih.gov/COG/).

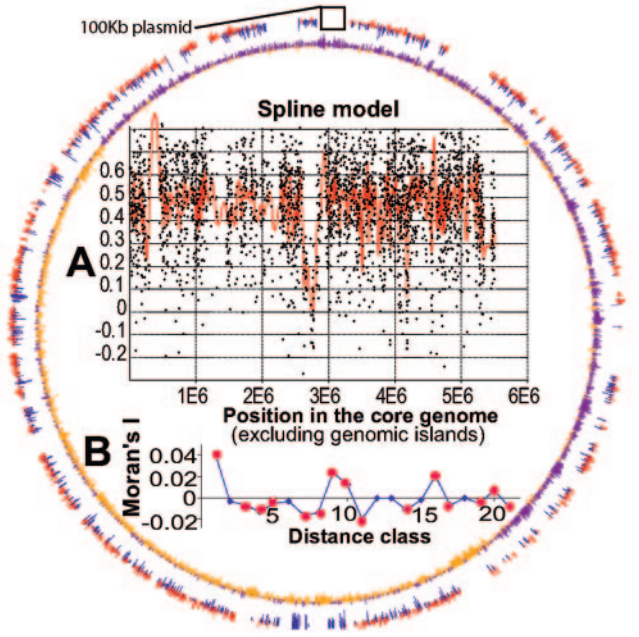

FIG. 6. Spatial distribution of the Kendall τ values for individual genes in the *E. coli* genome. The inner circle represents the G + C% skew analysis of the *E. coli* strain O157 (Sakai) genome, while the outer circle represents the distribution of the Kendall τ values for every core gene (n = 2,635), centered around the average Kendall τ value, which is ~0.46. Kendall τ values are derived from the correlation analysis between the individual-gene-based ML distances and the ANIo distances for all pairs of *E. coli* genomes (see Materials and Methods for details). Blue bars represent genes with Kendall τ values that are smaller than and red bars genes with Kendall τ values that are higher than the average Kendall τ value, while the height of the bar is proportional to the difference from the average Kendall τ value. The figure was plotted using the GenomeViz software (7). (A) The individual values were plotted along the mean position of each gene, and a local fitting algorithm was used to further reveal the major patterns in the data. (B) The presence of significant autocorrelation, as represented by filled circles, was determined for each distance class by a Moran *I* correlogram by bootstrap analysis (see Materials and Methods for more detail).

appear, however, that the functionality of the genes might have an effect on this periodicity. For instance, the major drop in Kendall τ values around the six o'clock position in the circle of the Sakai genome corresponds to that of the genes in flagella and pili operons. These genes are known to be among the most variable and most frequently horizontally transferred genes in the genome, and hence, it is not surprising that they showed poor Kendall τ correlation with ANIo.

DISCUSSION

Phylogenetic sequence analysis of multiple (six to eight) genes in the genome represents currently the most favorable approach for studying species diversity. Our evaluations of four important bacterial groups show that this approach is highly reliable for phylogenetic analysis and for discriminating among strains of different, even very closely related, species (using the current 70% DDH or 96% ANIo standard for species demarcation) because it gives results that approximate well the results from whole-genome comparisons. Furthermore, it appears that the performance of the approach may be independent of the genes employed (Fig. 2 and 4). The genes employed appear to be critical, however, when targeting shorter evolutionary scales, i.e., the intraspecies level, because the phylogenies of individual genes frequently correlate poorly with the whole-genome phylogeny at this level. The methodology developed here and the availability of a few genomic sequences can guide the selection of the best-performing genes for this evolutionary scale and thus substantially increase the robustness of the approach (Fig. 3). Our analyses also showed that the best-performing genes at this level might belong to any functional category, and in fact, that the informational genes may be less reliable markers for microevolution studies, as exemplified by the *E. coli* group (Fig. 5A).

Once the highly reliable markers have been identified, they can be used to robustly predict whole-genome-level relatedness (i.e., ANI) among a large collection of uncharacterized strains of a species. This has important applications for microbiology, including metagenomic surveys, e.g., in assigning genomic fragments such as fosmid or bacterial artificial chromosome clones to specific genotypes, especially for clones with identical or nearly identical rRNA genes or for clones that lack the commonly used, universal markers, such as the ribosomal proteins, DNA polymerases, and RecA and GyrB genes. In the latter case, metabolic or even hypothetical genes may constitute reliable markers (Fig. 5). Three, in our opinion, is the minimum number of genes to use in such MLST-like applications because if there is an unanticipated horizontal gene transfer (HGT) or recombination event in one of the genes in one or a few lineages, the phylogenetic signal conflicting with the remaining two genes will uncover the HGT event. Moreover, our analysis suggests that a random selection of six to eight genes would be expected to give a statistically significant prediction of whole-genome relatedness, even in the worst-case scenario, in which the genes employed are among the worst-performing ones (Fig. 3). Therefore, the MLST data available to date should be reasonably predictive of ANIo values if there are a few reference genomic sequences available for calibrating the equation between MLST and ANIo values (i.e., it is necessary to understand how conserved, at the sequence level, the genes used in MLST are relative to the genome average). This also has important applications for demarcating species based on the current standards, using the MLST approach as opposed to the cumbersome DDH method, since the ANIo values can be predicted from MLST data (and 96% ANIo corresponds tightly to 70% DDH) (10).

It also appears that even closely related groups may show significant differences in terms of which are the very best performing genes within each group (Fig. 5). For example, the orthologs of the genes that appeared to be the best within the *E. coli* group were typically not among the best genes for the *Salmonella* group (data not shown). Nonetheless, the best genes for *E. coli* typically showed quite strong (but not among the strongest) correlations with the ANIo for the *Salmonella* group. Therefore, extrapolations from one group to another group are possible but need to be done with caution. A good understanding of the extent of genetic diversity within each group and the level of evolutionary relatedness that is targeted within each group is essential for more-accu-

rate extrapolations due to the differences between shorter (i.e., within-species) and longer evolutionary scales (Fig. 2).

The reasons for the poor correlation of individual genes with whole-genome-based relatedness at the intraspecies level are diverse and could include a lack of time for selection to act upon sequence conservation, more-frequent horizontal gene transfer and recombination events within species, varied levels of gene divergence among lineages, and intragenomic recombination, etc. It is important to point out that since the distances between strains of the same species are smaller than the distances between strains of different species, the exact same rates or levels of the processes described above will have more-dramatic effects in the phylogeny of strains of the same species. Presumably, all these processes, in combination or alone, are affecting at least a fraction of the genes in the genome, and their relative importance is a subject for future investigations. It is tempting to speculate, however, that the lack of elapsed time as opposed to HGT and recombination may have a relatively greater impact at the whole-genome level. The fact that almost all genes in the genome yield good correlations with ANIo when more-distant genomes are included in the analysis, such as when the *E. coli* group is expanded by including the *Salmonella* genomes (data not shown), favors the greater impact of the lack-of-divergence-time hypothesis.

We obtained very comparable results for the *E. coli* group, in all aspects of our analyses, when this group was expanded with eight unpublished genomic sequences (analytical data not shown). For instance, the linear regressions of the Kendall τ values for all individual genes to ANIo values between the 12-genome (66 nonredundant pairs of genomes used in the correlation analysis of each gene) and the 20-genome (190 nonredundant pairs of genomes) data sets gave an r^2 value of 0.81 (Table 1). These results underscore the robustness of our methodology and suggest that the findings from a small number of genomic sequences may be universally applicable to more genomes within the same group.

The analysis performed here also showed that the ANI of the conserved genes in the genome (10) could be used as the reference standard for measuring genetic or evolutionary relatedness within species or between closely related species (Fig. 1). For measurement of relatedness, ANI is much easier to conceptualize and compute than whole-genome-based ML. The fact that ANI is a simple, robust, and pragmatic measurement for all bacteria and provides a very robust representation of their phylogenetic relationships at the species and probably up to the family level greatly magnifies its importance and potential for finer-scale systematic, diversity, and epidemiological studies.

ACKNOWLEDGMENTS

We thank The Institute for Genomic Research and the Sanger Center for permission to use preliminary sequence data.

This work was supported by the National Science Foundation (awards DEB0516252 and DEB-00755564), the DOE Genomics: GTL Program (for sequencing and the *Shewanella* Federation), and the Center for Microbial Ecology. K.T.K. is grateful to the Bouyoukos Fellowship Program for supporting his Ph.D. studies, and A.R. acknowledges a Swiss National Foundation Fellowship for a postdoctoral position in the laboratory of J.M.T.

REFERENCES

1. **Adiri, R. S., U. Gophna, and E. Z. Ron.** 2003. Multilocus sequence typing (MLST) of *Escherichia coli* O78 strains. FEMS Microbiol. Lett. **222:**199–203.
2. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
3. **Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork.** 2006. Toward automatic reconstruction of a highly resolved tree of life. Science **311:**1283–1287.
4. **Feil, E. J.** 2004. Small change: keeping pace with microevolution. Nat. Rev. Microbiol. **2:**483–495.
5. **Feil, E. J., B. C. Li, D. M. Aanensen, W. P. Hanage, and B. G. Spratt.** 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol. **186:**1518–1530.
6. **Gevers, D., F. M. Cohan, J. G. Lawrence, B. G. Spratt, T. Coenye, E. J. Feil, E. Stackebrandt, Y. Van de Peer, P. Vandamme, F. L. Thompson, and J. Swings.** 2005. Opinion: re-evaluating prokaryotic species. Nat. Rev. Microbiol. **3:**733–739.
7. **Ghai, R., T. Hain, and T. Chakraborty.** 2004. GenomeViz: visualizing microbial genomes. BMC Bioinformatics **5:**198.
8. **Hyma, K. E., D. W. Lacher, A. M. Nelson, A. C. Bumbaugh, J. M. Janda, N. A. Strockbine, V. B. Young, and T. S. Whittam.** 2005. Evolutionary genetics of a new pathogenic Escherichia species: *Escherichia albertii* and related *Shigella boydii* strains. J. Bacteriol. **187:**619–628.
9. **Kishino, H., and M. Hasegawa.** 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. **29:**170–179.
10. **Konstantinidis, K. T., and J. M. Tiedje.** 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. USA **102:**2567–2572.
11. **Konstantinidis, K. T., and J. M. Tiedje.** 2005. Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. **187:**6258–6264.
12. **Konstantinidis, K. T., and J. M. Tiedje.** 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc. Natl. Acad. Sci. USA **101:**3160–3165.
13. **Legendre, P., and L. Legendre.** 1998. Numerical ecology, 2nd English edition. Elsevier Science BV, Amsterdam, The Netherlands.
14. **Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. USA **95:**3140–3145.
15. **Nemoy, L. L., M. Kotetishvili, J. Tigno, A. Keefer-Norris, A. D. Harris, E. N. Perencevich, J. A. Johnson, D. Torpey, A. Sulakvelidze, J. G. Morris, Jr., and O. C. Stine.** 2005. Multilocus sequence typing versus pulsed-field gel electrophoresis for characterization of extended-spectrum beta-lactamase-producing *Escherichia coli* isolates. J. Clin. Microbiol. **43:**1776–1781.
16. **Posada, D., and T. R. Buckley.** 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst. Biol. **53:**793–808.
17. **Posada, D., and K. A. Crandall.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14:**817–818.
18. **SPSS Inc.** 1999. SPSS Base 10.0 for Windows user's guide. SPSS Inc., Chicago, Ill.
19. **Stackebrandt, E., W. Frederiksen, G. M. Garrity, P. A. D. Grimont, P. Kampfer, M. C. J. Maiden, X. Nesme, R. Rossello-Mora, J. Swings, H. G. Truper, L. Vauterin, A. C. Ward, and W. B. Whitman.** 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. **52:**1043–1047.
20. **Swofford, D. L.** 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.
21. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
22. **Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304:**66–74.
23. **Wayne, L. G., D. J. Brenner, R. R. Colwell, P. A. D. Grimont, O. Kandler, M. I. Krichevsky, L. H. Moore, W. E. C. Moore, R. G. E. Murray, E. Stackebrandt, M. P. Starr, and H. G. Trüper.** 1987. Report of the Ad Hoc Committee on reconciliation of approaches to Bacterial Systematics. Int. J. Syst. Bacteriol. **37:**463–464.